AD-754 751

MULTIWAY CONTINGENCY TABLE ANALYSIS APPLIED TO THE CLASSIFICATION OF MULTIVARIATE DICHOTOMOUS POPULATIONS

S. Kullback

George Washington University

Prepared for:

Office of Naval Research

9 January 1973

DISTRIBUTED BY:

NTIS

National Technical Information Service
U. S. DEPARTMENT OF COMMERCE
5285 Port Royal Road, Springfield Va. 22151

MULTIWAY CONTINGENCY TABLE ANALYSIS APPLIED TO THE CLASSIFICATION OF MULTIVARIATE DICHOTOMOUS POPULATIONS

by

S. KULLBACK

TECHNICAL REPORT NO. 4

January 9, 1973

PREPARED UNDER CONTRACT N00014-67-A-0214-0015

(NR-042-267)

OFFICE OF NAVAL RESEARCH

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, D. C. 20006

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) THE GEORGE WASHINGTON UNIVERSITY DEPT. OF STATISTICS WASHINGTON, D. C. 20006	2a. REPORT SECURITY CLASSIFICATION
	2b. GROUP

3. REPORT TITLE

MULTIWAY CONTINGENCY TABLE ANALYSIS APPLIED TO THE CLASSIFICATION OF MULTIVARIATE DICHOTOMOUS POPULATIONS

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

TECHNICAL REPORT

5. AUTHOR(S) (Last name, first name, initial)

KULLBACK, S.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
January 9, 1973	22	8

8a. CONTRACT OR GRANT NO. N00014-67-A-0214-0015	9a. ORIGINATOR'S REPORT NUMBER(S) #4
b. PROJECT NO. NR-042-267	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. AVAILABILITY/LIMITATION NOTICES

Unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Office of Naval Research Statistics & Probability Program Arlington, Va. 22217

13. ABSTRACT

Multiway contingency tables, or cross-classifications of vectors of discrete random variables, provide a useful approach to the analysis of multivariate discrete data. In the particular application we shall consider herein, the individual variates are dichotomous or binary. We shall use techniques and concepts presented and discussed by the author in previous papers. We note that the procedures and analysis are not restricted to dichotomous or binary data but are also applicable to polychotomous variates. The procedure we shall use is based on the principle of minimum discrimination information estimation applied to the analysis of multiway contingency tables. It yields results practically equivalent to procedures proposed by other investigators. When the minimum discrimination information estimates provide a satisfactory fit to a set of data, a complete analysis, including significance tests and estimates describing the pattern of observations is provided.

DD FORM 1 JAN 64 1473

Unclassified

Security Classification

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A ROLE	LINK A WT	LINK B ROLE	LINK B WT	LINK C ROLE	LINK C WT
CONTINGENCY TABLES						
DICHOTOMOUS POPULATIONS						
MULTIVARIATE						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through ________________."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through ________________."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through ________________."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.

DD FORM 1 JAN 64 1473 (BACK)

Unclassified

Security Classification

II

Multiway Contingency Table Analysis Applied to the Classification of Multivariate Dichotomous Populations

by

S. Kullback

Introduction

Multiway contingency tables, or cross-classifications of vectors of discrete random variables, provide a useful approach to the analysis of multivariate discrete data. In the particular application we shall consider herein, the individual variates are dichotomous or binary. We shall use techniques and concepts presented and discussed in [4] and [6]. We note that the procedures and analysis are not restricted to dichotomous or binary data but are also applicable to polychotomous variates.

For background on the study and problem which gave rise to the data we shall analyze see [8]. In [3], procedures further developed in [4] and [6], were applied to problems of multivariate binary data in information systems, such as communication, pattern recognition, and learning systems. In [1] there is a review of methods and models for the analysis of multivariate binary data. Solomon's data, which we shall analyze herein, is given as a typical example. In [7] there is developed a model based on a set of orthogonal polynomials and applied to Solomon's data. We remark that the procedure we shall use, based on the principle of minimum discrimination information estimation applied to the analysis of multiway contingency tables yields a result practically equivalent to that in [7].

"Multivariate data analysis needs a large and flexible class of hypothetical distributions of free variables indexed by the values of fixed variables. From this class, appropriate subfamilies would be chosen for fitting to specific data sets" [2]. The principle of minimum discrimination information estimation, and its basis, the minimum discrimination information theorem which is quite general in its formulation, lead to exponential families of distributions [4], [5], [6], The exponential families have very useful and desirable statistical properties and contain many subfamilies in common use [2]. "The data analytic attitude to models is empirical rather than theoretical... when detailed theoretical understanding is unavailable, a more empirical attitude is natural, so that estimation of parameters in models should be seen less as attempts to discover underlying truth and more as data calibrating devices which make it easier to conceive of noisy data in terms of smooth distributions and relations. Exponential families are viewed here as intended for use in the empirical mode. With a given data set, a variety of models may be tried on, and one selected on the ground of looks and fit" [2]. When the minimum discrimination information estimates provide a satisfactory fit to a set of data, a complete analysis, including significance tests and estimates describing the pattern of observations is provided.

Solomon's Data

A total of 2982 high-school seniors were given an attitude questionnaire to assess their attitude towards science. The students were also

classified on the basis of an IQ test into high IQ, the upper half, and low IQ, the lower half. The sixteen possible response vectors to each of four agree-disagree responses were tabulated. The data is given in table 1, where x_1, x_2, x_3, x_4 indicate the statements ([8,p.416]), agree and disagree were coded as 1 and 0 respectively, and listed as low IQ and high IQ. The problem of interest was to determine whether the response vectors could be used as a basis for classifying the students into one of two classes and evaluate possible classification procedures.

Contingency Table Analysis

We shall treat the data as a five-way $2\times2\times2\times2\times2$ contingency table, denoting the original observations by $x(hijk\ell)$, where

$h=1$, low IQ, $h=2$, high IQ ;

$i=1$, response to x_1 coded 0, $i=2$, response to x_1 coded 1;

$j=1$, response to x_2 coded 0, $j=2$, response to x_2 coded 1;

$k=1$, response to x_2 coded 0, $k=2$, response to x_3 coded 1;

$\ell=1$, response to x_4 coded 0, $\ell=2$, response to x_4 coded 1.

As a first overview of the data to determine the marginals and their related interaction parameters which may be considered to furnish significant values in the log-linear representation of the exponential family of the estimates [6], we list in table 2a, Analysis of Information, a sequential study of interaction and effect type measures [4], [6].

We remark that the first estimate is

$$x_a^*(hijk\ell) = x(h\cdot\cdot\cdot\cdot)x(\cdot ijk\ell)/n$$

and the minimum discrimination information statistic (interaction type measure)

$$2I(x:x_a^*) = 2\Sigma\Sigma\Sigma\Sigma\Sigma\ x(hijk\ell)\ell n\ \frac{x(hijk\ell)n}{x(h\cdots\cdot)x(\cdot ijk\ell)}$$

tests a null hypothesis that the IQ groupings are homogeneous over the sixteen response vectors [5,Chap.8], [4]. This null hypothesis is rejected and the subsequent study of effect and interaction type measures is an attempt to get a good fit to the data and account for the variation. Although the association between IQ and the response to the first statement is not significant, $2I(x_b^*:x_a^*) = 2.376$, 1 D.F., it was decided to examine in detail the estimate $x_e^*(hijk\ell)$ whose numerical values are given in table 1. (We remark that it may be shown that

$$2I(x_b^*:x_a^*) = 2\Sigma\Sigma\ x(hi\cdots)\ell n\ \frac{x(hi\cdots)n}{x(h\cdots\cdot)x(\cdot i\cdots)}$$

and tests a null hypothesis that IQ is homogeneous over the response to the first question). The estimate $x_e^*(hijk\ell)$ was selected because it does not differ significantly from the observed values, $2I(x:x_e^*) = 16.307$, 11D.F. (represents an acceptable fit), is symmetric with respect to the four statements, and is comparable to the first-order model estimate of [7], whose values are also listed in table 1.

From the log-linear representation in figure 1 [6], we obtain the parametric representation for the log-odds (low IQ/high IQ)

$$\ell n(x_e^*(1ijk\ell)/x_e^*(2ijk\ell))$$

over the sixteen response vectors as given in table 3a. Thus, for example

$$\ln \frac{x_e^*(11111)}{x_e^*(21111)} = \tau_1^h + \tau_{11}^{hi} + \tau_{11}^{hj} + \tau_{11}^{hk} + \tau_{11}^{h\ell} ,$$

that is, a linear regression of the log-odds in terms of a constant τ_1^h and the main effects of each component of the response vector, namely, τ_{11}^{hi}, τ_{11}^{hj}, τ_{11}^{hk}, $\tau_{11}^{h\ell}$. The numerical values of the log-odds and the parameters are easily obtained from the entries in the computer output and are also given in table 3a [6].

We note from table 3a that

$$\ln \frac{x_e^*(1ijk1)}{x_e^*(2ijk1)} - \ln \frac{x_e^*(1ijk2)}{x_e^*(2ijk2)} = \tau_{11}^{h\ell} = 0.3338 ,$$

that is, a change from disagree to agree on the fourth statement is associated with an increase of 0.3338 in the log-odds (low IQ/high IQ). Note also that $\tau_{11}^{h\ell}$ represents the association between IQ and response to the fourth statement as measured by the log-cross-product - ratio

$$\tau_{11}^{h\ell} = \ln \frac{x_e^*(1ijk1)x_e^*(2ijk2)}{x_e^*(2ijk1)x_e^*(1ijk2)}$$

and is the same for all eight levels of the responses to statements one, two and three.

Similarly, it is found that

$$\ln \frac{x_e^*(1ij1\ell)}{x_e^*(2ij1\ell)} - \ln \frac{x_e^*(1ij2\ell)}{x_e^*(2ij2\ell)} = \tau_{11}^{hk} = 0.3411 ,$$

$$\ln \frac{x_e^*(1i1k\ell)}{x_e^*(2i1k\ell)} - \ln \frac{x_e^*(1i2k\ell)}{x_e^*(2i2k\ell)} = \tau_{11}^{hj} = 0.1240 ,$$

$$\ln \frac{x_e^*(11jk\ell)}{x_e^*(21jk\ell)} - \ln \frac{x_e^*(12jk\ell)}{x_e^*(22jk\ell)} = \tau_{11}^{hi} = -0.2030 .$$

Classification

Since $x(1\cdot\cdot\cdot\cdot) = x^*_e(1\cdot\cdot\cdot\cdot) = 1491$, and $x(2\cdot\cdot\cdot\cdot) = x^*_e(2\cdot\cdot\cdot\cdot) = 1491$, we assign a response vector $(ijk\ell)$ to the region

E_1: classify as population $h=1$ (low IQ), when

$$\ell n \frac{x^*_e(1ijk\ell)}{x^*_e(2ijk\ell)} \geqq 0$$

and to the complementary region

E_2: classify as population $h=2$ (high IQ), when

$$\ell n \frac{x^*_e(1ijk\ell)}{x^*_e(2ijk\ell)} < 0 \ .$$

If we set

$$\mu_1(E_1) = \sum_{(ijk\ell)\epsilon E_1} \frac{x^*_e(1ijk\ell)}{1491}, \quad \mu_2(E_1) = \sum_{(ijk\ell)\epsilon E_1} \frac{x^*_e(2ijk\ell)}{1491} \ ,$$

then the probability of error of the classification procedure is [5,pp.4,69,80]

$$\text{Prob Error} = p\mu_2(E_1)+q\mu_1(E_2) = (\mu_2(E_1)+\mu_1(E_2))/2$$

since here $p = x(2\cdot\cdot\cdot\cdot)/2982 = \frac{1}{2}$, $q = x(1\cdot\cdot\cdot\cdot)/2982 = \frac{1}{2}$.

The relevant computations with $x^*_e(hijk\ell)$ are given in table 4(b) and show that the Prob. Error = 0.444. The corresponding computations with the original data $x(hikj\ell)$ are given in table 4(a) and yield Prob. Error = 0.441.

Other Estimates

In view of the measure of the effect of the marginal $x(hi\cdot\cdot\ell)$ (and the associated interaction parameters) in table 2a, $2I(x^*_m:x^*_g) = 4.316$, 1D.F.

and the marginal $x(h \cdot j \cdot \ell)$, $2I(x^*_p : x^*_n) = 3.181$, 1 D.F., the estimate $x^*_v(hijk\ell)$ fitting the marginals $x(\cdot ijk\ell)$, $x(h \cdot j \cdot \cdot)$, $x(h \cdot \cdot k \cdot)$, $x(hi \cdot \cdot \ell)$ and the estimate $x^*_w(hijk\ell)$ fitting the marginals $x(\cdot ijk\ell)$, $x(h \cdot \cdot k \cdot)$, $x(hi \cdot \cdot \ell)$, $x(h \cdot j \cdot \ell)$ were computed. The estimates are given in table 1 and the relevant analysis of information given in table 2b.

The values of the log-odds, parametric representation, and the values of the associated interaction parameters are given in table 3b for $x^*_v(hijk\ell)$ and in table 3c for $x^*_w(hijk\ell)$. Note from table 3b that

$$\ell n \frac{x^*_v(11jk1)}{x^*_v(21jk1)} - \ell n \frac{x^*_v(11jk2)}{x^*_v(21jk2)} = \tau^{h\ell}_{11} + \tau^{hi\ell}_{111} = 0.6469 \ ,$$

$$\ell n \frac{x^*_v(12jk1)}{x^*_v(22jk1)} - \ell n \frac{x^*_v(12jk2)}{x^*_v(22jk2)} = \tau^{h\ell}_{11} = 0.2680 \ ,$$

$$\ell n \frac{x^*_v(11jk1)}{x^*_v(21jk1)} - \ell n \frac{x^*_v(12jk1)}{x^*_v(22jk1)} = \tau^{hi}_{11} + \tau^{hi\ell}_{111} = -0.0276$$

$$\ell n \frac{x^*_v(11jk2)}{x^*_v(21jk2)} - \ell n \frac{x^*_v(12jk2)}{x^*_v(22jk2)} = \tau^{hi}_{11} = -0.4065$$

reflecting the interaction of the responses to the first and fourth statements.

From table 3c, it is found for example, that

$$\ell n \frac{x^*_w(111k1)}{x^*_w(211k1)} - \ell n \frac{x^*_w(111k2)}{x^*_w(211k2)} = \tau^{h\ell}_{11} + \tau^{hi\ell}_{111} + \tau^{hj\ell}_{111} = 0.5806$$

$$\ell n \frac{x^*_w(121k1)}{x^*_w(221k1)} - \ell n \frac{x^*_w(121k2)}{x^*_w(221k2)} = \tau^{h\ell}_{11} + \tau^{hj\ell}_{111} = 0.2030$$

$$\ell n \frac{x^*_w(112k1)}{x^*_w(212k1)} - \ell n \frac{x^*_w(112k2)}{x^*_w(212k2)} = \tau^{h\ell}_{11} + \tau^{hi\ell}_{111} = 0.9371$$

$$\ell n \frac{x^*_w(122k1)}{x^*_w(222k1)} - \ell n \frac{x^*_w(122k2)}{x^*_w(222k2)} = \tau^{h\ell}_{11} = 0.5595$$

reflecting the interactions of the responses to the first, second and fourth statements.

The computation of the probability of error using the estimates $x^*_v(hijk\ell)$ and $x^*_w(hijk\ell)$ is shown in table 4(c) and 4(d) respectively, and yields probabilities of error 0.444 and 0.446.

Measure of Divergence

As a measure of the divergence between the low IQ and high IQ observed and estimated values, we computed the values of

$$J(1,2) = \frac{1}{2} \Sigma\Sigma\Sigma\Sigma(x(1ijk\ell)-x(2ijk\ell))\ell n \frac{x(1ijk\ell)}{x(2ijk\ell)}$$

for $x(hijk\ell)$, $x^*_e(hijk\ell)$, $x^*_v(hijk\ell)$, $x^*_w(hijk\ell)$ [5,p.130]. The resulting values and their ratios to the respective degrees of freedom are given in table 5. As is to be expected from the properties of the discrimination information we note that

$$J(1,2;x^*_e) < J(1,2;x^*_v) < J(1,2;x^*_w) < J(1,2;x) .$$

However the ratio to the respective degrees of freedom leads to the inequalities

$$J(1,2;x)/D.F. < J(1,2;x^*_e)/D.F. < J(1,2;x^*_v)/D.F. < J(1,2;x^*_w)/D.F.$$

Remark

Martin and Bradley [7] examined Solomon's data in terms of an estimate they called a first-order or linear model. These estimated values are

given in table 1. It turns out that although the underlying approaches are different, the Martin and Bradley parameters and estimates arc practically the same as those for $x^*_e(hijk\ell)$. From [7,pp.216-217] we note that

$$\ell n \frac{x^*_e(12222)}{x^*_e(22222)} = \tau_1^h = \ell n \frac{1+a_o+a_1+a_2+a_3+a_4}{1-a_o-a_1-a_2-a_3-a_4}$$

$$\ell n \frac{x^*_e(12221)}{x^*_e(22221)} = \tau_1^h + \tau_{11}^{h\ell} = \ell n \frac{1+a_o+a_1+a_2+a_3-a_4}{1-a_o-a_1-a_2-a_3+a_4}$$

$$\ell n \frac{x^*_e(12212)}{x^*_e(22212)} = \tau_1^h + \tau_{11}^{hk} = \ell n \frac{1+a_o+a_1+a_2-a_3+a_4}{1-a_o-a_1-a_2+a_3-a_4}$$

$$\ell n \frac{x^*_e(12122)}{x^*_e(22122)} = \tau_1^h + \tau_{11}^{hj} = \ell n \frac{1+a_o+a_1-a_2+a_3+a_4}{1-a_o-a_1+a_2-a_3-a_4}$$

$$\ell n \frac{x^*_e(11222)}{x^*_e(21222)} = \tau_1^h + \tau_{11}^{hi} = \ell n \frac{1+a_o-a_1+a_2+a_3+a_4}{1-a_o+a_1-a_2-a_3-a_4}$$

or to a first approximation

$$\tau_1^h = 2a_o+2a_1+2a_2+2a_3+2a_4$$

$$\tau_1^h+\tau_{11}^{h\ell} = 2a_o+2a_1+2a_2+2a_3-2a_4$$

$$\tau_1^h+\tau_{11}^{hk} = 2a_o+2a_1+2a_2-2a_3+2a_4$$

$$\tau_1^h+\tau_{11}^{hj} \doteq 2a_o+2a_1-2a_2+2a_3+2a_4$$

$$\tau_1^h+\tau_{11}^{hi} = 2a_o-2a_1+2a_2+2a_3+2a_4 .$$

It is found that

$$\tau_{11}^{h\ell} = -4a_4$$

$$\tau_{11}^{hk} = -4a_3$$

$$\tau_{11}^{hj} = -4a_2$$

$$\tau_{11}^{hi} = -4a_1 \ .$$

The values of the parameters given in [7, table 3, p. 217] are

$$a_0 = -0.042, \quad a_1 = 0.049, \quad a_2 = -0.031, \quad a_3 = -0.084, \quad a_4 = -0.082$$

so that

$$\tau_{11}^{h\ell} = 0.3338 = 0.334, \quad -4a_4 = 0.328$$

$$\tau_{11}^{hk} = 0.3411 = 0.341, \quad -4a_3 = 0.336$$

$$\tau_{11}^{hj} = 0.1240 = 0.124, \quad -4a_2 = 0.124$$

$$\tau_{11}^{hi} = -0.2030 = -0.203, \quad -4a_1 = -0.196$$

The computation for the probability of error using the estimates in [7] are shown in table 4(e) and yields a probability of error 0.445. (Martin and Bradley give a value of the risk as 0.455).

Acknowledgment

This report was prepared under Navy Contract N0014-67-A0214-0015 and partially supported by the Air Force Office of Scientific Research, Office of Aerospace Research. U.S. Air Force under Grant AFOSR-72-2348.

h i j k l	1 L	2 h 1	3 i 1	4 j 1	5 k 1	6 l 1	7 hi 11	8 hj 11	9 hk 11	10 hl 11	11 ij 11	12 ik 11	13 il 11	14 jk 11	15 jl 11	16 kl 11	17 hij 111	18 hik 111	19 hil 111	20 hjk 111	21 hjl 111	22 hkl 111	23 ijk 111	24 ijl 111	25 ikl 111	26 jkl 111	27 hijk	28 hijl	29 hikl	30 hjkl	31 ijkl	32 hijkl
1 1 1 1 1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1 1 1 1 2	1	1	1	1	1		1	1	1		1	1		1			1	1		1			1				1					
1 1 1 2 1	1	1	1	1		1	1	1		1	1		1		1		1		1		1			1				1				
1 1 1 2 2	1	1	1	1			1	1			1						1															
1 1 2 1 1	1	1	1		1	1	1		1	1		1	1			1		1	1			1			1				1			
1 1 2 1 2	1	1	1		1		1		1			1						1														
1 1 2 2 1	1	1	1			1	1			1			1						1													
1 1 2 2 2	1	1	1				1																									
1 2 1 1 1	1	1		1	1	1		1	1	1				1	1	1				1	1	1				1				1		
1 2 1 1 2	1	1		1	1			1	1					1						1												
1 2 1 2 1	1	1		1		1		1		1					1						1											
1 2 1 2 2	1	1		1				1																								
1 2 2 1 1	1	1			1	1			1	1						1						1										
1 2 2 1 2	1	1			1				1																							
1 2 2 2 1	1	1				1				1																						
1 2 2 2 2	1	1																														
2 1 1 1 1	1		1	1	1	1					1	1	1	1	1	1							1	1	1	1					1	
2 1 1 1 2	1		1	1	1						1	1		1									1									
2 1 1 2 1	1		1	1		1					1		1		1									1								
2 1 1 2 2	1		1	1							1																					
2 1 2 1 1	1		1		1	1						1	1			1									1							
2 1 2 1 2	1		1		1							1																				
2 1 2 2 1	1		1			1							1																			
2 1 2 2 2	1		1																													
2 2 1 1 1	1			1	1	1								1	1	1										1						
2 2 1 1 2	1			1	1									1																		
2 2 1 2 1	1			1		1									1																	
2 2 1 2 2	1			1																												
2 2 2 1 1	1				1	1										1																
2 2 2 1 2	1				1																											
2 2 2 2 1	1					1																										
2 2 2 2 2	1																															
x		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
x^*_e		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓							✓	✓	✓	✓					✓	
x^*_v		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓				✓	✓	✓	✓					✓	
x^*_w		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓		✓		✓	✓	✓	✓					✓	

Figure 1

Solomon's Data-Classification Procedures

$x_1x_2x_3x_4$	ij kℓ	Observed Low IQ $x(11jk\ell)$	Martin & Bradley	Estimates $x_e^*(11jk\ell)$	$x_v^*(11jk\ell)$	$x_w^*(11jk\ell)$	Observed High IQ $x(21jk\ell)$	Martin & Bradley	Estimates $x_e^*(21jk\ell)$	$x_v^*(21jk\ell)$	$x_w^*(21jk\ell)$
11 11	22 22	62	74.56	74.589	76.097	70.156	122	109.45	109.414	107.904	113.844
11 10	22 21	70	67.30	67.296	66.198	71.600	68	70.71	70.703	71.802	66.400
11 01	22 12	31	31.32	31.329	31.943	29.827	33	32.68	32.671	32.057	34.173
11 00	22 11	41	37.74	37.780	37.337	39.884	25	28.26	28.219	28.662	26.115
10 11	21 22	283	266.76	266.570	271.120	275.979	329	345.24	345.429	340.879	336.020
10 10	21 21	253	259.17	259.322	254.876	250.769	247	240.83	240.679	245.125	249.232
10 01	21 12	200	193.45	193.625	196.841	200.037	172	178.55	178.376	175.160	171.963
10 00	21 11	305	314.50	314.491	310.589	306.748	217	207.50	207.508	211.411	215.252
01 11	12 22	14	12.10	12.156	10.866	9.914	20	21.90	21.844	23.135	24.085
01 10	12 21	11	9.20	9.182	9.929	10.760	10	11.80	11.818	11.071	10.240
01 01	12 12	11	9.68	9.659	8.776	8.102	11	12.32	12.341	13.224	13.898
01 00	12 11	14	12.02	12.010	12.855	12.756	9	10.98	10.990	10.144	9.244
00 11	11 22	31	33.63	33.623	30.125	30.820	56	53.37	53.375	56.874	56.179
00 10	11 21	46	47.37	47.263	50.789	50.001	55	53.63	53.737	50.211	50.999
00 01	11 12	37	47.54	47.450	43.233	44.163	64	53.46	53.550	57.767	56.837
00 00	11 11	82	74.67	74.656	79.426	78.482	53	60.33	60.346	55.574	56.517
		1491					1491				

Table 1

Table 2a

Analysis of Information

Marginals Fitted	Information	D.F.
a) x(.ijkℓ),x(h....)	$2I(x:x^*_a) = 68.369$	15
b) x(.ijkℓ),x(hi...)	$2I(x^*_b:x^*_a) = 2.376$	1
	$2I(x:x^*_b) = 65.993$	14
c) x(.ijkℓ),x(hi...),x(h.j..)	$2I(x^*_c:x^*_b) = 4.265$	1
	$2I(x:x^*_c) = 61.728$	13
d) x(.ijkℓ),x(hi...),x(h.j..),x(h..k.)	$2I(x^*_d:x^*_c) = 25.230$	1
	$2I(x:x^*_d) = 36.498$	12
e) x(.ijkℓ),x(hi...),x(h.j..),x(h..k.),x(h...ℓ)	$2I(x^*_e:x^*_d) = 20.191$	1
	$2I(x:x^*_e) = 16.307$	11
f) x(.ijkℓ),x(h..k.),x(h...ℓ),x(hij..)	$2I(x^*_f:x^*_e) = 3.016$	1
	$2I(x:x^*_f) = 13.291$	10
g) x(.ijkℓ),x(h...ℓ),x(hij..),x(hi.k.)	$2I(x^*_g:x^*_f) = 0.042$	1
	$2I(x:x^*_g) = 13.249$	9
m) x(.ijkℓ),x(hij..),x(hi.k.),x(hi..ℓ)	$2I(x^*_m:x^*_g) = 4.316$	1
	$2I(x:x^*_m) = 8.933$	8
n) x(.ijkℓ),x(hij..),x(hi.k.),x(hi..ℓ),x(h.jk.)	$2I(x^*_n:x^*_m) = 0.983$	1
	$2I(x:x^*_n) = 7.950$	7
p) x(.ijkℓ),x(hij..),x(hi.k.),x(hi..ℓ),x(h.jk.),x(h.j.ℓ)	$2I(x^*_p:x^*_n) = 3.181$	1
	$2I(x:x^*_p) = 4.769$	6
q) x(.ijklℓ),x(hij..),x(hi.k.),x(hi..ℓ),x(h.jk.),x(h.j.ℓ), x(h..kℓ)	$2I(x^*_q:x^*_p) = 0.219$	1
	$2I(x:x^*_q) = 4.550$	5
r) x(.ijkℓ),x(hi..ℓ),x(h.j.ℓ),x(h..kℓ),x(hijk.)	$2I(x^*_r:x^*_q) = 0.346$	1
	$2I(x:x^*_r) = 4.204$	4

Analysis of Information (continued)

Marginals Fitted	Information	D.F.
	$2I(x:x_r^*) = 4.204$	4
s) $x(\cdot ijk\ell), x(h\cdot\cdot k\ell), x(hijk\cdot), x(hij\cdot\ell)$	$2I(x_s^*:x_r^*) = 2.303$	1
	$2I(x:x_s^*) = 1.901$	3
t) $x(\cdot ijk\ell), x(hijk\cdot), x(hij\cdot\ell), x(hi\cdot k\ell)$	$2I(x_t^*:x_s^*) = 1.375$	1
	$2I(x:x_t^*) = 0.526$	2
u) $x(\cdot ijk\ell), x(hijk\cdot), x(hij\cdot\ell), x(hi\cdot k\ell), x(h\cdot jk\ell)$	$2I(x_u^*:x_t^*) = 0.361$	1
	$2I(x:x_u^*) = 0.165$	1

Table 2b

Analysis of Information

Marginals Fitted	Information	D.F.
e) $x(\cdot ijk\ell), x(hi\cdot\cdot\cdot), x(h\cdot j\cdot\cdot), x(h\cdot\cdot k\cdot), x(h\cdot\cdot\cdot\ell)$	$2I(x:x_e^*) = 16.307$	11
v) $x(\cdot ijk\ell), x(h\cdot j\cdot\cdot), x(h\cdot\cdot k\cdot), x(hi\cdot\cdot\ell)$	$2I(x_v^*:x_e^*) = 3.735$	1
	$2I(x:x_v^*) = 12.572$	10
w) $x(\cdot ijk\ell), x(h\cdot\cdot k\cdot), x(hi\cdot\cdot\ell), x(h\cdot j\cdot\ell)$	$2I(x_w^*:x_v^*) = 3.443$	1
	$2I(x:x_w^*) = 9.129$	9

$$\text{Log-odds} \quad \ell n \frac{x_e^*(1ijk\ell)}{x_e^*(2ijk\ell)}$$

ijkℓ	Parametric representation					log-odds
1111	τ_1^h	$+\tau_{11}^{hi}$	$+\tau_{11}^{hj}$	$+\tau_{11}^{hk}$	$+\tau_{11}^{h\ell}$	0.2128
1112	τ_1^h	$+\tau_{11}^{hi}$	$+\tau_{11}^{hj}$	$+\tau_{11}^{hk}$		-0.1210
1121	τ_1^h	$+\tau_{11}^{hi}$	$+\tau_{11}^{hj}$		$+\tau_{11}^{h\ell}$	-0.1284
1122	τ_1^h	$+\tau_{11}^{hi}$	$+\tau_{11}^{hj}$			-0.4621
1211	τ_1^h	$+\tau_{11}^{hi}$		$+\tau_{11}^{hk}$	$+\tau_{11}^{h\ell}$	0.0888
1212	τ_1^h	$+\tau_{11}^{hi}$		$+\tau_{11}^{hk}$		-0.2450
1221	τ_1^h	$+\tau_{11}^{hi}$			$+\tau_{11}^{h\ell}$	-0.2524
1222	τ_1^h	$+\tau_{11}^{hi}$				-0.5861
2111	τ_1^h		$+\tau_{11}^{hj}$	$+\tau_{11}^{hk}$	$+\tau_{11}^{h\ell}$	0.4158
2112	τ_1^h		$+\tau_{11}^{hj}$	$+\tau_{11}^{hk}$		0.0820
2121	τ_1^h		$+\tau_{11}^{hj}$		$+\tau_{11}^{h\ell}$	0.0746
2122	τ_1^h		$+\tau_{11}^{hj}$			-0.2592
2211	τ_1^h			$+\tau_{11}^{hk}$	$+\tau_{11}^{h\ell}$	0.2918
2212	τ_1^h			$+\tau_{11}^{hk}$		-0.0420
2221	τ_1^h				$+\tau_{11}^{h\ell}$	-0.0494
2222	τ_1^h					-0.3831

$\tau_1^h = -0.3831$, $\tau_{11}^{hi} = -0.2030$, $\tau_{11}^{hj} = 0.1240$

$\tau_{11}^{hk} = 0.3411$, $\tau_{11}^{h\ell} = 0.3338$

Table 3a

Log-odds $\ell n \dfrac{x_v^*(1ijk\ell)}{x_v^*(2ijk\ell)}$

ijkℓ	Parametric representation						log-odds
1111	τ_1^h	$+\tau_{11}^{hi}$	$+\tau_{11}^{hj}$	$+\tau_{11}^{hk}$	$+\tau_{11}^{h\ell}$	$+\tau_{111}^{hi\ell}$	0.3571
1112	τ_1^h	$+\tau_{11}^{hi}$	$+\tau_{11}^{hj}$	$+\tau_{11}^{hk}$			-0.2898
1121	τ_1^h	$+\tau_{11}^{hi}$	$+\tau_{11}^{hj}$		$+\tau_{11}^{h\ell}$	$+\tau_{111}^{hi\ell}$	0.0115
1122	τ_1^h	$+\tau_{11}^{hi}$	$+\tau_{11}^{hj}$				-0.6355
1211	τ_1^h	$+\tau_{11}^{hi}$		$+\tau_{11}^{hk}$	$+\tau_{11}^{h\ell}$	$+\tau_{111}^{hi\ell}$	0.2366
1212	τ_1^h	$+\tau_{11}^{hi}$		$+\tau_{11}^{hk}$			-0.4101
1221	τ_1^h	$+\tau_{11}^{hi}$			$+\tau_{11}^{h\ell}$	$+\tau_{111}^{hi\ell}$	-0.1088
1222	τ_1^h	$+\tau_{11}^{hi}$					-0.7557
2111	τ_1^h		$+\tau_{11}^{hj}$	$+\tau_{11}^{hk}$	$+\tau_{11}^{h\ell}$		0.3847
2112	τ_1^h		$+\tau_{11}^{hj}$	$+\tau_{11}^{hk}$			0.1167
2121	τ_1^h		$+\tau_{11}^{hj}$		$+\tau_{11}^{h\ell}$		0.0390
2122	τ_1^h		$+\tau_{11}^{hj}$				-0.2290
2211	τ_1^h			$+\tau_{11}^{hk}$	$+\tau_{11}^{h\ell}$		0.2644
2212	τ_1^h			$+\tau_{11}^{hk}$			-0.0036
2221	τ_1^h				$+\tau_{11}^{h\ell}$		-0.0813
2222	τ_1^h						-0.3492

$$\tau_1^h = -0.3492, \quad \tau_{11}^{hi} = -0.4065, \quad \tau_{11}^{hj} = 0.1203$$

$$\tau_{11}^{hk} = 0.3457, \quad \tau_{11}^{h\ell} = 0.2680, \quad \tau_{111}^{hi\ell} = 0.3789$$

Table 3b

Log-odds $\ell n \dfrac{x^*_w(1ijk\ell)}{x^*_w(2ijk\ell)}$

$ijk\ell$	Parametric representation							log-odds
1111	τ^h_1	$+\tau^{hi}_{11}$	$+\tau^{hj}_{11}$	$+\tau^{hk}_{11}$	$+\tau^{h\ell}_{11}$	$+\tau^{hi\ell}_{111}$	$+\tau^{hj\ell}_{111}$	0.3283
1112	τ^h_1	$+\tau^{hi}_{11}$	$+\tau^{hj}_{11}$	$+\tau^{hk}_{11}$				-0.2523
1121	τ^h_1	$+\tau^{hi}_{11}$	$+\tau^{hj}_{11}$		$+\tau^{h\ell}_{11}$	$+\tau^{hi\ell}_{111}$	$+\tau^{hj\ell}_{111}$	-0.0197
1122	τ^h_1	$+\tau^{hi}_{11}$	$+\tau^{hj}_{11}$					-0.6004
1211	τ^h_1	$+\tau^{hi}_{11}$		$+\tau^{hk}_{11}$	$+\tau^{h\ell}_{11}$	$+\tau^{hi\ell}_{111}$		0.3976
1212	τ^h_1	$+\tau^{hi}_{11}$		$+\tau^{hk}_{11}$				0.5396
1221	τ^h_1	$+\tau^{hi}_{11}$			$+\tau^{h\ell}_{11}$	$+\tau^{hi\ell}_{111}$		0.0495
1222	τ^h_1	$+\tau^{hi}_{11}$						-0.8876
2111	τ^h_1		$+\tau^{hj}_{11}$	$+\tau^{hk}_{11}$	$+\tau^{h\ell}_{11}$		$+\tau^{hj\ell}_{111}$	0.3542
2112	τ^h_1		$+\tau^{hj}_{11}$	$+\tau^{hk}_{11}$				0.1512
2121	τ^h_1		$+\tau^{hj}_{11}$		$+\tau^{h\ell}_{11}$		$+\tau^{hj\ell}_{111}$	0.0061
2122	τ^h_1		$+\tau^{hj}_{11}$					-0.1968
2211	τ^h_1			$+\tau^{hk}_{11}$	$+\tau^{h\ell}_{11}$			0.4235
2212	τ^h_1			$+\tau^{hk}_{11}$				-0.1360
2221	τ^h_1				$+\tau^{h\ell}_{11}$			0.0754
2222	τ^h_1							-0.4841

$\tau^h_1 = -0.4841,\ \tau^{hi}_{11} = -0.4035,\ \tau^{hj}_{11} = 0.2873$

$\tau^{hk}_{11} = 0.3481,\ \tau^{h\ell}_{11} = 0.5595,\ \tau^{hi\ell}_{111} = 0.3776$

$\tau^{hj\ell}_{111} = -0.3565$

Table 3c

E_1: $\{ijk\ell: \ell n \text{ odds} \geq 0\}$

E_1: Observations

ijkℓ	$x(1ijk\ell)$	$x(2ijk\ell)$
1111	82	53
1211	14	9
1221	11	10
2111	305	217
2112	200	172
2121	253	247
2211	41	25
2221	70	68
	976	801

$$\mu_2(E_1) = \frac{801}{1491}, \quad \mu_1(E_2) = \frac{1491-976}{1491}$$

$$\text{Prob. Error} = \frac{1}{2}\,\frac{801+515}{1491} = \frac{1316}{2\text{x}1491} = 0.441$$

(a)

E_1: x^*_e

ijkℓ	$x^*_e(1ijk\ell)$	$x^*_e(2ijk\ell)$
1111	74.656	60.346
1211	12.010	10.990
2111	314.491	207.508
2112	193.625	178.376
2121	259.322	240.679
2211	37.780	28.219
	891.884	726.118

$$\mu_2(E_1) = \frac{726.118}{1491}, \quad \mu_1(E_2) = \frac{1491-891.884}{1491}$$

$$\text{Prob. Error} = \frac{1}{2}\,\frac{726.118+599.116}{1491} = \frac{1325.234}{2982} = 0.444$$

(b)

Table 4

$E_1 : x_v^*$

ijkℓ	$x_v^*(11jk\ell)$	$x_v^*(21jk\ell)$
1111	79.426	55.574
1121	50.789	50.211
1211	12.855	10.144
2111	310.589	211.411
2112	196.841	175.160
2121	254.876	245.125
2211	37.337	28.662
	942.713	776.287

$$\mu_2(E_1) = \frac{776.287}{1491}$$

$$\mu_1(E_2) = \frac{1491-942.713}{1491}$$

$$\text{Prob. Error} = \frac{1}{2}\ \frac{776.287+548.287}{1491}$$

$$= \frac{1324.574}{2982}$$

$$= 0.444$$

(c)

$E_1 : x_w^*$

ijkℓ	$x_w^*(11jk\ell)$	$x_w^*(21jk\ell)$
1111	78.482	56.517
1211	13.756	9.244
1212	8.102	13.898
1221	10.760	10.240
2111	306.748	215.252
2112	200.037	171.963
2121	250.769	249.232
2211	39.884	26.115
2221	71.600	66.401
	980.138	818.862

$$\mu_2(E_1) = \frac{818.862}{1491}$$

$$\mu_1(E_2) = \frac{1491-980.138}{1491}$$

$$\text{Prob. Error} = \frac{1}{2}\ \frac{818.862+510.862}{1491}$$

$$= \frac{1329.724}{2982}$$

$$= 0.446$$

(d)

Table 4

E_1	$\hat{x}(1ijk\ell)$	$\hat{x}(2ijk\ell)$
1111	74.67	60.33
1211	12.02	10.98
2111	314.50	207.50
2112	193.45	178.55
2121	259.17	240.83
2211	37.74	28.26
	891.55	726.45

$$\mu_2(E_1) = \frac{726.45}{1491}\,, \quad \mu_1(E_2) = \frac{1491-891.55}{1491}$$

$$\text{Prob. Error} = \frac{1}{2}\ \frac{726.45+599.45}{1491}$$

$$= \frac{1325.90}{2982}$$

$$= 0.445$$

Table 4(e)

Divergence Between Low IQ and High IQ

Observations and Estimates

$$\frac{1}{2}\Sigma\Sigma\Sigma\Sigma(x(1ijk\ell)-x(2ijk\ell))\ell n\,\frac{x(1ijk\ell)}{x(2ijk\ell)} = 69.132$$

$$69.132/15 = 4.61/\text{D.F.}$$

$$\frac{1}{2}\Sigma\Sigma\Sigma\Sigma(x^*_e(1ijk\ell)-x^*_e(2ijk1))\ell n\,\frac{x^*_e(1ijk\ell)}{x^*_e(2ijk\ell)} = 52.374$$

$$52.374/11 = 4.76/\text{D.F.}$$

$$\frac{1}{2}\Sigma\Sigma\Sigma\Sigma(x^*_v(1ijk\ell)-x^*_v(2ijk\ell))\ell n\,\frac{x^*_v(1ijk\ell)}{x^*_v(2ijk\ell)} = 56.249$$

$$56.249/10 = 5.62/\text{D.F.}$$

$$\frac{1}{2}\Sigma\Sigma\Sigma\Sigma(x^*_w(1ijk\ell)-x^*_w(2ijk\ell))\ell n\,\frac{x^*_w(1ijk\ell)}{x^*_w(2ijk\ell)} = 59.815$$

$$59.815/9 = 6.65/\text{D.F.}$$

Table 5

References

[1] Cox, D.R. (1972), The analysis of multivariate binary data, Applied Statistics, 21, 113-120.

[2] Dempster, A.P. (1971), An overview of multivariate data analysis, Journal of Multivariate Analysis, 1, 316-346.

[3] Ku, H.H. and Kullback, S. (1969), Approximating discrete probability distributions, IEEE Trans. on Information Theory, IT-15, 444-447.

[4] Ku, H.H., Varner, R.N., and Kullback, S. (1971), On the analysis of multidimensional contingency tables, Journal of the American Statistical Association, 66, 55-64.

[5] Kullback, S. (1959), Information Theory and Statistics, Wiley, N.Y. 1968 Edition, Dover Publications Inc. N.Y.

[6] Kullback, S. (1970), Minimum discrimination information estimation and application, Invited paper presented to Sixteenth Conference on the Design of Experiments in Army Research, Development and Testing, 21 October 1970. ARO-D Report 71-3, 1-38 Proceedings of the Conference.

[7] Martin, D.C. and Bradley, R.A. (1972), Probability models, estimation, and classification for multivariate dichotomous populations, Biometrics, 28, 203-221.

[8] Solomon, H. (1960), Classification procedures based on dichotomous response vectors, No. 36 in Contributions to Probability and Statistics, Essays in Honor of Harold Hotelling, Edited by I. Olkin et al Stanford U.P., Stanford, Cal. 1960, pp.414-423 (Also in Studies in Item Analysis and Prediction, Edited by H. Solomon, Stanford U.P. Stanford, Cal, 1961, pp. 177-186.